Random Variables and Probability Functions

Stephen R. Addison

February 9, 2001

Basic Definitions

Given a random variable x, we develop a probability distribution function and define it mean value, variance, and standard deviation. Let each value of x, x_i have a probability of occurrence p_i. We write $p_i = f(x_i) =$ probability that $x = x_i$. The function $f(x)$ is variously known as the probability function, the frequency function, or the probability distribution function for the random variable x. For now we will limit x and $f(x)$ to a finite number of discrete values. (This restriction isn't necessary, but it simplifies things.)

Now we can define a random variable. A variable x is random if it takes on the values x_i with probabilities $p_i = f(x_i)$. Random variables are called random simply because they take on a variety of values. A random process is one whose outcome isn't known in advance. Random processes are also called stochastic.

Mean Value and Standard Deviation

$f(x)$ provides complete information, we will often need less. Consider N measurements (of length for instance) of x, x_i where N is large.

$$p_i = f(x_i) \propto N_i$$

where N_i is the number of times x_i was obtained, i.e.

$$p_i = \frac{N_i}{N}.$$

To average, we sum and divide by the number of measurements in the usual way. Thus

$$\overline{x} = <x> = \frac{1}{N}\sum_i N_i x_i = \sum_i \frac{N_i}{N} x_i = \sum_i p_i x_i$$

$\overline{x}$, $<x>$ deonote the sample mean, the true mean is denoted by μ.

For any probability distribution function $f(x)$, the average value of x is

$$\overline{x} = \sum_i x_i f(x_i),$$

where x is any random variable. In quantum mechanics such averages are called expectation values. We can develop distribution functions from probability distribution functions. Let

$$F(x_i) = p(x \leq x_i) = \sum_{x_j \leq x_i} f(x_i),$$

$F(x)$, the distribution function, is then the curve of the sum of the $f(x_j)$.

Variance and Standard Deviation

Both measure the spread of data values around the mean, if s is the variance and σ is the standard deviation, then

$$\sigma(x) = \sqrt{s(x)}.$$

If we know the true mean

$$s(x) = \sum_i (x_i - \mu)^2 f(x_i).$$

If we only know $\overline{x}$, the sample mean, then

$$s(x) = \frac{N}{N-1} \sum_i (x_i - \overline{x})^2 f(x_i).$$

This latter definition prevents calculation of the spread in one value, σ and s become closer to their true values as N increases.

Continuous Distributions

When we let x take on a continuous set of values, we need to replace sums by integrals. The idea of an integral as the limit of a sum is a familiar (in fact a defining) idea. Let $dN(x)$ be the number of elements that have values of x between x and $x + dx$, then we can define f(x) by

$$dN(x) = Np(x)dx = Nf(x)dx.$$

There are N total elements so

$$\int dN(x) = \int_R Nf(x)dx = N$$

in other words

$$\int_R f(x)dx = 1$$

R is the range over which x can vary. An alternate statement is to say that when we look at the system, then the chance of it being between x and $x + dx$

is $f(x)dx$.
The average value of x,

$$<x>= \int_{-\infty}^{\infty} xf(x)dx$$

where I've replaced the range R by $\pm\infty$, assuming $f(x)$ is zero if the probability is zero.
Similarly

$$\overline{x^2} = \int_{-\infty}^{\infty} x^2 f(x)dx$$

and

$$s(x) = \int_{-\infty}^{\infty} (x-\mu)^2 = \sigma^2(x)$$

We will usually need the mean.

The Binomial Distribution

Consider coin tossing, what is the probability of getting 3 heads in 5 tosses. The probability of a sequence such as *ththt* is $\left(\frac{1}{2}\right)^5$, since the tosses are independent. We know how to count the number of such sequences so let's approach this problem directly. In other words how many ways are there of choosing 3 heads from 5 coins, we don't care about the result for particular coins so there are $C(5,3)$ sequences. The probability of getting 3 heads in 5 tosses is

$$p(3 \text{ heads}|5 \text{ tosses}) = C(5,3)\left(\frac{1}{2}\right)^5$$

We can view this as

$$\frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{C(5,3)}{2^5}$$

We generalize this to find the probability function $p = f(x)$ that gives the probability of x heads in n tosses.

$$f(x) = \frac{C(5,3)}{2^n}$$

What's the probability of 3 ones in 5 rolls of a die?
$A = 1$ $N = \text{not}1$

$$p(A) = \frac{1}{6}$$

$$p(N) = \frac{5}{6}$$

$$p(AANNA) = \frac{1}{6}\frac{1}{6}\frac{5}{6}\frac{5}{6}\frac{1}{6}$$

The number of sequences with 3 A's and 2 N's is $C(5,3)$. So the probability of 3 ones and 2 other outcomes is

$$f(x) = C(5,3)\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2$$

Where we get $f(x)$ by multiplying the probability for a given sequence by the number of such sequences. This is equivalent to our usual approach.
A generalization $p(x$ ones in n rolls) is given by

$$f(x) = C(n,x)\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{n-x}$$

Convince yourselves that this is equivalent — try four ones out of five or something like that. In many problems we'll be interested in repeated independent trials where there are two independent outcomes in each trial.

Binomial Probability Functions

We've looked at a variety of problems where something is tried repeatedly and there are two outcomes with probabilities p and q.

$$p = \text{probability of success}$$

$$q = 1 - p = \text{probability of failure}$$

Such repeated independent trials — trials with constant probabilities — are known as Bernoulli trials. A general formula that gives the probability of n successes in n trials is

$$f(x) = C(n,x)p^x q^{n-x},$$

this formula is called the binomial probability function or the binomial distribution. What's the probability of not more than x successes in n trials, it's the sum of $0, 1, 2, 3, \ldots, x$ successes. That is

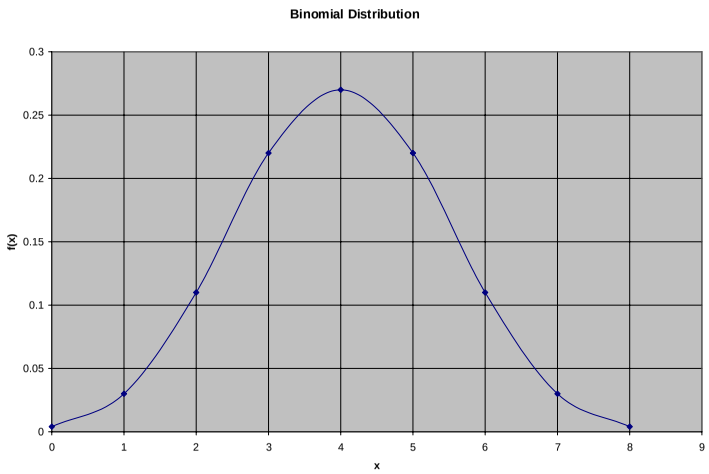
$$\begin{aligned} F(x) &= f(0) + f(1) + f(2) + f(3) + \ldots + f(x) \\ &= \sum_{i=0}^{x} C(n,i)p^i q^{n-i} \end{aligned}$$

In these equations the $f(i)$ represent the terms of $(p+q)^n$, hence the name. $F(x)$ is the binomial distribution function. Plots of $f(x)$ versus x are symmetric if $p = q$.

Calculating the binomial distribution is tedious unless n is small, we usually use approximations. (If both n and np are large, we get the Gaussian distribution, if $p \neq q$ we get the Poisson distribution when p is small.)

Consider $n = 8$, $p = \frac{1}{2}$ $f(x) = C(n,x)p^x q^{n-x}$

$$f(0) = (8,0)\left(\frac{1}{2}\right)^8 = 0.004$$



$$f(1) = C(8,1)\left(\frac{1}{2}\right)^8 = 0.03$$

$$f(2) = (8,2)\left(\frac{1}{2}\right)^8 = 0.11$$

$$f(3) = (8,3)\left(\frac{1}{2}\right)^8 = 0.22$$

$$f(4) = (8,4)\left(\frac{1}{2}\right)^8 = 0.27$$

$$f(5) = (8,5)\left(\frac{1}{2}\right)^8 = 0.22$$

$$f(6) = (8,6)\left(\frac{1}{2}\right)^8 = 0.11$$

$$f(7) = (8,7)\left(\frac{1}{2}\right)^8 = 0.03$$

$$f(8) = (8,8)\left(\frac{1}{2}\right)^8 = 0.004$$

The Normal or Gaussian Distribution

For large $n,\ np$ and $f(x) = C(n,x)p^x q^{n-x}$

$$\lim_{n\to\infty} \frac{f(x)}{(2\pi npq)^{-\frac{1}{2}} e^{-(x-np)^2/2npq}} = 1$$

or

$$f(x) \sim \frac{1}{\sqrt{2\pi npq}}\, e^{-\frac{(x-np)^2}{2npq}} \quad (n \to \infty)$$

We arrive at this result by using the Stirling approximation and some other approximations.

Stirling Approximation

For large N,

$$N! \sim (2\pi N)^{1/2} N^N e^{-N} \quad (N \to \infty)$$

The logarithm of this result is of more interest.

$$\begin{aligned}
\ln N! &= \ln(2\pi N)^{1/2} N^N e^{-N} \\
&= N \ln N - N \ln e + \frac{1}{2} \ln(2\pi N) \\
&= N \ln N - N + \frac{1}{2} \ln N + \frac{1}{2} \ln 2\pi \\
&\approx N \ln N - N
\end{aligned}$$

Probability that x lies in a particular range

The probability that x lies between x_1 and x_2 is

$$p(x_1 \le x \le x_2) \sim \frac{1}{\sqrt{2\pi npq}} \int_{x_1}^{x_2} e^{-\frac{(x-np)^2}{2npq}}\, dx$$

This is the area under the curve between x_1 and x_2.